MONTE CARLO SIMULATIONS OF THE NESTED FIXED-POINT ALGORITHM

Erik P. Johnson

Working Paper # WP2011-011
October 2010

http://www.econ.gatech.edu/research/wokingpapers

School of Economics
Georgia Institute of Technology
221 Bobby Dodd Way
Atlanta, GA 30332–0615

Monte Carlo Simulations of the Nested Fixed-Point Algorithm

Erik P. Johnson

Working Paper # WP2011-11
October 2010
JEL No. C18, C63

ABSTRACT

There have been substantial advances in dynamic structural models and in the econometric literature about techniques to estimate those models over the past two decades. One area in which these new developments has lagged is in studying robustness to distributional assumptions and finite sample properties in small samples. This paper extends our understanding of the behavior of these estimation techniques by replicating John Rust's (1987) influential paper using the nested fixed-point algorithm (NFXP) and then using Monte Carlo techniques to examine the finite sample properties of the estimator. I then examine the consequences of the distributional assumptions needed to estimate the model on the parameter estimates. I find that even in sample sizes of up to 8,000 observations, the NFXP can display finite sample bias and variances substantially larger than the theoretical asymptotic variance. This is also true with departures from distributional assumptions, with the mean square error increasing by a factor of 10 for some distributions of unobserved variables.

Erik P. Johnson
Georgia Institute of Technology
221 Bobby Dodd Way, Atlanta, GA 30332
erik.johnson@econ.gatech.edu

Monte Carlo Simulations of the Nested Fixed-Point Algorithm

Erik Johnson

October 2010

Preliminary and incomplete. Please do not cite or quote.

Abstract

There have been substantial advances in dynamic structural models and in the econometric literature about techniques to estimate those models over the past two decades. One area in which these new developments has lagged is in studying robustness to distributional assumptions and finite sample properties in small samples. This paper extends our understanding of the behavior of these estimation techniques by replicating John Rust's (1987) influential paper using the nested fixed-point algorithm (NFXP) and then using Monte Carlo techniques to examine the finite sample properties of the estimator. I then examine the consequences of the distributional assumptions needed to estimate the model on the parameter estimates. I find that even in sample sizes of up to 8,000 observations, the NFXP can display finite sample bias and variances substantially larger than the theoretical asymptotic variance. This is also true with departures from distributional assumptions, with the mean square error increasing by a factor of 10 for some distributions of unobserved variables.

1 Introduction

Structural estimation of dynamic programming problems has become a prominent tool in many economists' toolbox since the publication of John Rust's "The Optimal Replacement of GMC Bus Engines" [9], particularly in the context of dynamic discrete choice models[1]. Papers that use structural estimation are generally characterized by a complete, explicit, usually dynamic, mathematical model of agents' behavior, then estimate the parameters of the model either through maximum likelihood or method of moments. However, these models often rely on assumptions about the distributions of unobservables and functional forms to make them tractable to estimate. Even with these assumptions, they usually result in highly nonlinear objective functions that present a challenge to estimate. There is a growing literature that examines the how numerical methods such as choice of optimization routines and starting points affect estimates. For example, Knittel and Metaxoglou [6] investigate how researchers' decisions about maximization algorithms and different starting points for each algorithm can lead to different answers. They find a very wide array of estimates can be obtained depending on the choices a researcher makes about the maximization algorithm and starting points. However, there are no papers to my knowledge that examine how distributional assumptions affect parameter estimates. This paper attempts to begin to fill this hole by examining a very simple dynamic structural model.

This paper uses the optimal stopping model from Rust [9] of GMC bus engine replacement as a starting point to evaluate how distributional assumptions affect the performance of the nested fixed point algorithm (NFXP). Rust develops a dynamic discrete choice model of bus engine replacement for the supervisor of the Madison Metro Bus Company, Harold Zurcher. In each period

[1]For a survey of the dynamic discrete choice literature see Rust [11], Pakes [8], Miller [7], and Mira [2].

Mr. Zurcher observes the mileage that the bus has accumulated and has a discrete choice to make: replace the engine in the bus or use the current engine for another month. Rust then poses a functional form for Mr. Zurcher's utility function over bus engine replacements and assumes that Mr. Zurcher is a forward looking agent who dynamically maximizes this utility function. Rust then solves Mr. Zurcher's dynamic problem and finds parameter values that maximize the likelihood of the data. This involves solving the entire dynamic maximization problem for every set of parameter values. This is done through the nested fixed point algorithm. The nested fixed point algorithm is an inner loop that solves a dynamic programming problem for a given set of parameter values and an outer loop that uses a routine to to maximize the likelihood function over the parameter space.

Since Rust developed this framework for solving dynamic discrete choice problems, there have been many algorithms proposed to solve similar problems. Hotz and Miller [4] showed that it is not necessary to solve the dynamic problem at every step like the nested fixed point algorithm requires. Instead, since there is a one to one mapping between conditional choice probabilities and normalized value functions, the conditional probabilities can be inverted into estimates of the value functions which in turn allow the econometrician to update the conditional probabilities. Aguirregabiria and Mira [1] show that the nested fixed point algorithm and Hotz and Miller's conditional choice probabilities estimator are two extreme cases of a general class of estimators.

There has generally been seen to be a tradeoff between efficiency (from the nested fixed point algorithm) and computation time (reduced by using the Hotz and Miller [4] routine). I have chosen to use the NFXP from Rust [9] as a starting point for this paper for two reasons. Firstly, since this was one of the first papers to employ a structural approach to a dynamic problem it has become one of the standards in evaluating new methods. This is partly because the algorithm is particularly

easy to implement. For instance, Hotz et. al [5] perform Monte Carlo simulations to compare their conditional choice simulation estimator to the NFXP and examine the NFXP for sample sizes of 10,000 and more. Secondly, because the NFXP solves the dynamic programming problem at every step I expect that it would be more robust to specification error.

This paper contributes to the literature on structural estimation in two ways. First, it extends the range of sample sizes for which there is Monte Carlo evidence for the validity of the nested fixed point algorithm and similar algorithms. In this paper I simulate the NFXP for datasets with as few as 500 observations and as many as 11,800. (Previously, the literature had only examined sample sizes as small as 10,000 [5].) Given that many papers that use structural estimation of a dynamic programming problem rely on sample sizes less than 10,000 I feel this is the relevant range of observations. Second, I examine how distributional assumptions on the unobserved state variable effect the estimates of the structural parameters. While this is obviously a context specific effect it is still important to have a sense of how important these assumptions may be in parameter estimates.

The remainder of this paper is organized as follows. Section 2 describes in detail Rust's model of bus engine replacement. Section 3 describes the data that is used in Rust [9]. Section 4 discusses Rust's results and my replication of his results. Section 5 discusses asymptotic results for the estimation procedure and how I simulate the data. Section 6 discusses the simulation results and Section 7 concludes.

2 Rust's Model

Rust provides two versions of his model. The first model, which I name the simple model, imposes strict functional form assumptions on the transition probabilities and assumes that there are no unobserved state variables. The second model, which I call the relaxed model, relaxes the functional form assumption on the transition probabilities and introduces an unobserved (to the econometrician) state variable, ϵ_t, that Rust assumes has very specific properties.

2.1 The Simple Model

John Rust [9] models the behavior of the superintendent of the Madison Wisconsin Metropolitan Bus Company, Harold Zurcher, when deciding whether or not to replace the engine in one of the company's buses. The model takes the form of a regenerative optimal stopping problem. Each month, Mr. Zurcher must choose either to (i) leave the bus in service for another month, while doing "normal maintenance" and incur operating costs $c(x_t, \theta_1)$ or (ii) take the bus out of service for the month and completely replace the engine for a cost of $\bar{P}$ and sell the old engine for scrap for a price of $\underline{P}$. (Let the replacement cost of the engine, $RC = \bar{P} - \underline{P}$.) Mr. Zurcher is assumed to be a rational actor who minimizes the expected discounted costs of maintaining the fleet of buses. It is assumed that a bus with a newly replaced engine is just as good as a new bus in terms of the future decisions of whether or not to replace the engine. Therefore, the optimal stopping problem takes the form:

$$V_\theta(x_t) = \sup_{\Pi} E\left\{ \sum_{j=t}^{\infty} \beta^{j-t} u(x_j, f_j, \theta_1) \middle| x_t \right\} \tag{1}$$

where

$$u(x_t, i_t, \theta_1) = \begin{cases} -c(x_t, \theta_1) & \text{if } i_t = 0 \\ -(RC + c(0, \theta_1)) & \text{if } i_t = 1 \end{cases} \tag{2}$$

where Π is an infinite sequence of decision rules $\Pi = f_t, f_{t+1}, \dots$ where each f_t specifies Mr. Zurcher replacement decision at time t as a function of the entire history of the process, $i_t = f(x_t, i_{t-1}, x_{t-1}, i_{t-2}, \dots)$ and the expectation is taken with respect to the controlled stochastic process, $\{x_t\}$ whose probability distribution is defined from Π and the transition probability $p(x_{t+1}|x_t, i_t, \theta_2)$. If an exponential distribution is assumed for $p(x_{t+1}|x_t, i_t, \theta_2)$ then the transition probabilities take the form,

$$p(x_{t+1}|x_t, i_t, \theta_2) = \begin{cases} \theta_2 \exp[-\theta_2(x_{t+1} - x_t)] & \text{if } i_t = 0 \text{ and } x_{t+1} \geq x_t \\ \theta_2 \exp[-\theta_2(x_{t+1})] & \text{if } i_t = 1 \text{ and } x_{t+1} \geq 0 \end{cases} \tag{3}$$

Therefore, if the current engine is kept $(i_t = 0)$ the next period's mileage is given by a draw from the exponential distribution $1 - \exp[-\theta_2(x_{t+1} - x_t)]$, but if the engine is replaced $(i_t = 1)$ then x_t regenerates to 0 and the next period's mileage is drawn from the exponential distribution $1 - \exp[-\theta_2(x_{t+1} - 0)]$.

I can write the Bellman's equation to this system as:

$$V_\theta(x_t) = \max_{i_t \in \{0,1\}} [u(x_t, i_t, \theta_1) + \beta E V_\theta(x_{t+1}, i_{t+1})] \tag{4}$$

This should imply a deterministic cut-off rule such that

$$i_t = f(x_t, \theta) = \begin{cases} 1 & \text{if } x_t > \gamma(\theta_1, \theta_2) \\ 0 & \text{if } x_t \leq \gamma(\theta_1, \theta_2) \end{cases} \tag{5}$$

for some function $\gamma(\cdot)$.

However since in the data, we do not observe this type of deterministic cut-off rule, we assume that there is an unobserved state variable, ϵ_t, that Mr. Zurcher observes but the econometrician does not observe.

2.2 The Relaxed Model

Rust now adds two parts to the model. We add the unobserved state variable, ϵ_t, which is assumed to be additively separable from the rest of the utility function. Also, he relaxes the assumption the the mileage is drawn from an exponential distribution with parameter θ_2 and allow the mileage process to have an arbitrary density and define the difference between this month's mileage and last month's mileage to have arbitrary density $g(\cdot)$. These new assumptions lead to the Bellman's equation:

$$V_\theta(x_t, \epsilon_t) = \max_{i_t \in \{0,1\}} [u(x_t, i_t, \theta_1) + \epsilon_t(i) + \beta E V_\theta(x_{t+1}, \epsilon_{t+1})] \tag{6}$$

which has the solution

$$f(x_t, \epsilon_t, \theta) = \arg \max_{i_t \in \{0,1\}} [u(x_t, i, \theta_1) + \epsilon_t(i) + \beta E V_\theta(x_{t+1}, \epsilon_{t+1})] \tag{7}$$

Because the unobserved state variable, ϵ_t, enters non-linearly into the unknown function, EV_θ Rust makes a "Conditional Independence" assumption to circumvent this problem. The conditional independence assumption can be stated as:

Assumption 1 *Conditional Independence: The transition density of the controlled process* $\{x_t, \epsilon_t\}$ *factors as*

$$p(x_{t+1}, \epsilon_{t+1} | x_t, \epsilon_t, i, \theta_2, \theta_3) = q(\epsilon_{t+1} | x_{t+1}, \theta_2) p(x_{t+1} | x_t, i, \theta_3)$$

This assumption introduces two restrictions. First it requires that x_{t+1} is a sufficient statistic for ϵ_{t+1}, which means that any dependence between ϵ_t and ϵ_{t+1} is transmitted through x_{t+1}. Secondly, it requires that the probability density of x_{t+1} depends only on x_t and not ϵ_t. [2]

If we further impose that $q(\epsilon|y, \theta_2)$ is given by a type 1 extreme value distribution then we can state the formula for the choice probability, $P(i|x, \theta)$ as follows:

$$P(i|x,\theta) = \frac{\exp[u(x,i,\theta_1) + \beta EV_\theta(x,i)]}{\sum_{j\in\{0,1\}} \exp[u(x,j,\theta_1) + \beta EV_\theta(x,j)} \tag{8}$$

which is the familiar multinomial logit formula.

This allows us to estimate the structural parameters, $\theta \equiv \{RC, \theta_1, \theta_3\}$, of the controlled process $\{i_t, x_t\}$ through maximum likelihood as shown in Rust [10]. The likelihood function ℓ^f take the form

$$\ell^f(x_1,\ldots,x_T,i_1,\ldots,i_T|x_0,i_0,\theta) = \prod_{t=1}^{T} P(i_t|x_t,\theta)p(x_t|x_{t-1},i_{t-1},\theta_3) \tag{9}$$

This likelihood function can be estimated in three stages. The first stage is to estimate

$$\ell^1(x_1,\ldots,x_T,i_1,\ldots,i_T|x_0,i_0,\theta) = \prod_{t=1}^{T} p(x_t|x_{t-1},i_{t-1},\theta_3) \tag{10}$$

which is the transition probabilities between mileage bins. The second stage is to estimate

$$\ell^2(x_1,\ldots,x_T,i_1,\ldots,i_T|x_0,i_0,\theta) = \prod_{t=1}^{T} P(i_t|x_t,\theta) \tag{11}$$

which requires the computation of the fixed point to get estimates of θ_1 and RC, the variable cost parameter and the replacement cost of the engine respectively. Since estimating both ℓ^1 and ℓ^2 give consistent estimates of the parameters, I can then use these consistent estimates to estimate ℓ^f and

[2] For proofs of these results see Rust [10].

get efficient estimates of all of the structural parameters.

3 Data

I have obtained the relevant parts of Rust's original data from the Madison Metropolitan Bus Company that contains monthly maintenance records for every bus in the Madison bus fleet from December, 1974 to May, 1985.[3] The observations consist of odometer readings on each bus and an indicator specifying if the engine was replaced that month. In addition to the original data that I have obtained, Rust's original data also consisted of a maintenance diary that records all repairs that were made on a bus such as replacing brakes, oil changes, etc. Rust considers all events that are not a complete engine replacement "normal maintenance" and disregards that information for the sake of his exercise. I proceed likewise.

The data that I have from Rust contains the mileage for each bus at the end of every month, an indicator if the bus' engine was replaced in that month, and the model of the bus. There are eight types of buses in the Madison Metro fleet over the covered time period. See Tables 1a and 1b for summary statistics of the data.

In order to compute the value function in the dynamic programming problem, I will need to do a grid search. This means that I will need to discretize our mileage data into bins. I discretize the continuous mileage variable into 90 bins of 5,000 miles each.[4] This gives bins up to 450,000 miles to allow for that value function to be estimated for mileages above what I observe in the data (the maximum mileage I observe is 387,300) and allows for the possibility that it may be optimal

[3] The data used in the original paper is available at http://gemini.econ.umd.edu/jrust/nfxp.html

[4] Rust [9] does some sensitivity analysis by increasing the number of mileage bins and finds results similar to those with 90 bins.

to replace the engine at a mileage level large than I observe.

Now that I have discretized the mileage process, I can rewrite the transition density as the difference between last month's bin and this month's bin, giving density

$$p(x_{t+1}|x_t, i_t, \theta_3) = \begin{cases} g(x_{t+1} - x_t, \theta_3) & \text{if } i_t = 0 \\ g(x_{t+1} - 0, \theta_3) & \text{if } i_t = 1 \end{cases} \tag{12}$$

In the data I only have buses where $(x_{t+1} - x_t) \in \{0, 1, 2\}$, thus I define θ_{30} as the probability that you stay in the same mileage bin as you were last month, θ_{31}, as the probability that you move to the next mileage bin, and θ_{32} as the probability that you move up two mileage bins. This reduces to a multinomial distribution with parameters $\theta_{30},\ \theta_{31}$ (and $\theta_{32} = 1 - \theta_{30} - \theta_{31}$).

In order to use the nested fixed point algorithm I need to assume that there is no heterogeneity in our data between the different types of buses. Rust tests the hypothesis that the mileage process is different for various groupings of bus types and cannot reject the null that bus types 1-4 have the same mileage process, while you can reject the null that bus types 1-4 have a different mileage process from types 5-8.[5] Therefore, I proceed with the exercise only using bus types 1-4.

Next, I need to specify a functional form for the cost function. Rust did not find one particular functional form to fit the data statistically better than any other that he tried and therefore used a linear cost function[6] with one unknown parameter defined as $c(x, \theta_1) = .001\theta_{11}x$.

Following Rust, I choose to fix β instead of estimating it since it is highly collinear with the fixed cost of replacement, RC. This collinearity becomes obvious by examining the value function since a lower discount factor will weight the present higher, which has the same effect as raising RC. Following Rust, for the rest of the paper I fix $\beta = 0.9999$.

[5] See Rust [9] for a detailed analysis and results.

[6] A square root cost function, $c(x, \theta_1) = \theta_{11}\sqrt{x}$ was also used, but results not reported.

4 Rust's Results and Replication

Using the exact data that Rust [9] uses, I proceed with the replication using Rust's methods, detailed in Rust [12][7]. Rust [9] uses the nested fixed point algorithm to solve Mr. Zurcher's dynamic discrete control problem. This algorithm consists of two loops. The inside loop uses a combination of two methods to compute the fixed point. The first method used is the commonly used is value function contraction iterations. Value function iteration defines a fixed point as $EV_\theta = T_\theta(EV_\theta)$ where $T_\theta(W)$ is defined as

$$T_\theta(W)(x) = \int_0^\infty log[\exp\{-c(x+y),\theta) + \beta W(x+y)\} + \exp\{-RC - c(0,\theta) + \beta W(0)\}]g(dy|\theta)$$

This method begins with an arbitrary guess for EV_θ (usually equal to zero) and evaluates the value function given parameters θ and iterates the process k times. The k^{th} iteration can be written as $EV_k = T_\theta^k(EV_0)$ and as $k \to \infty$ it can be shown that $EV_k \to EV_\theta$.

Value function iteration converges at a linear rate to EV_θ. An alternative method, known as Newton-Kantorovich iteration uses an alternate method of iteration that converges at a quadratic rate when in the neighborhood of EV_θ. Thus, I use Werner's method [13] which uses value function iteration for the first few contraction steps and then switches to the Newton-Kantorovich method. Werner [13] showed that this produces a faster rate of convergence than either method alone.

Once the value function has converged, I evaluate the log-likelihood function using the assumed parameters $\theta_{11}, \theta_{30}, \theta_{31}, RC$. To get a new guess for the structural parameter, I use the outer hill climbing algorithm to find the parameters that maximize the likelihood function. Following Rust, I

[7]The replication was done using similar GAUSS code to that available through John Rust's website (http://gemini.econ.umd.edu/jrust/nfxp.html)

use the BHHH algorithm, which is similar to the Gauss-Newton and Newton-Raphson algorithms.

The results are presented in Table 2. As can be seen, my estimates of the transition probabilities are nearly identical, though the estimates of the cost function parameter, θ_{11}, and the replacement cost, RC, differ somewhat. It seems likely that I have found a slightly different local maximum than the original paper. Ideally, I would start the NFXP routine at many starting values and compare the value of the likelihood function at all maximum that the algorithm converges to in order to choose the global maximum. However, since for this paper it is only important that the routine always find "the same" maximum I will always initialize the algorithm to the same starting values so that it will likely head to the same local maximum.

5 Asymptotic Results and Simulation Procedure

Rust [10] shows that parameters estimated using the nested fixed-point maximum likelihood(NFXP) algorithm, $\hat{\theta}$, converges to the true value, θ^* with probability 1 as either N, the number of observations, or T, the number of periods, approaches infinity. He also shows that $\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, -H(\theta^*)^{-1})$ where $-H(\theta^*)^{-1}$ is the negative inverse of the Hessian for θ^*. The main assumptions needed to make this result hold is that the model is correctly specified, the Conditional Independence assumption:

$$p(x_{t+1}, \epsilon_{t+1}|x_t, \epsilon_t, i, \theta_2, \theta_3) = q(\epsilon_{t+1}|x_{t+1}, \theta_2)p(x_{t+1}|x_t, i, \theta_3),$$

and some regularity conditions[8].

Having validated the results reported in Rust [9], I examine the finite sample properties of the

[8]For a complete proof and set of assumptions see Theorem 4.3 of Rust [10].

maximum likelihood estimator using the nested fixed point algorithm. To do this, I simulate 1,000 datasets that are generated assuming that the model posed by Rust is correct and use the values that I estimated in the replication section as the "true" values of the model.

In order to simulate bus replacement data, there are two levels of randomness that need to be incorporated. First, the mileage bin that a bus falls into in a given month is a random variable that I model with a multinomial distribution, the parameters of the distribution, $\{\theta_{30}, \theta_{31}\}$ are random variables. Secondly, the model assumes an unobserved state variable, ϵ, that enters additively into the utility function and is independent across time and choices that is drawn from a Type I extreme value distribution.

In order to perform the actual simulation I need to proceed in a chronological order for each bus. Each bus is assumed to have an odometer reading of zero at the beginning of the simulation. In the first period the each bus receives a draw from the multinomial distribution for which mileage bin it will end that period in.

Once I know which mileage bin each bus ended the period in, I then evaluate the solution to the Bellman's equation, given parameters, θ.

$$f(x_t, \epsilon_t, \theta) = \arg \max_{i_t \in \{0,1\}} [u(x_t, i, \theta_1) + \epsilon_t(i) + \beta EV_\theta(x_{t+1}, \epsilon_{t+1})] \tag{13}$$

where

$$u(x_t, i, \theta_1) = \begin{cases} -[0.001\theta_{11} x_t + \epsilon_t(0)] & \text{if } i_t = 0 \\ -[RC + 0.001\theta_{11} x_0 + \epsilon_t(1)] & \text{if } i_t = 1 \end{cases} \tag{14}$$

and $\epsilon_t(\cdot)$ is drawn from a Type I extreme value distribution with mean 0 and variance $\frac{\pi^2}{6}$. If the value of replacement is larger than the value of not replacing the engine, then $i_t = 1$ and the bus starts over at mileage bin zero the next period. Once I have done this simulation for one month,

I repeat the process for each of the 110 buses in each dataset for 118 months[9]. This leaves each dataset with approximately 13,000 bus-month observations.

6 Simulation Results

I first report the results from the simulations using datasets with relatively small sample sizes and then will discuss the results from simulations where the data generating process (DGP) is not the assumed DGP in the model. I find two largely consistent themes across all of the simulations. First, the estimator is biased in all samples examined, for all 4 parameters, though the bias decreases as I get closer to the assumptions of the model (unobservables become closer in distribution to the assumed EV1 unobservables). Second, the asymptotic variance is substantially smaller than the observed variance of the distribution of parameters.

6.1 "Small" Sample Results

Using the procedure described above, I produced datasets with 1,000, 2,000, 4,000, and 8,000 bus-month observations. Knittel and Metaxoglou [6] have shown that the choice of starting values can create very different results in highly non-linear environments. Therefore, in all of the simulations I use the same starting value, which is within 0.1 of the true value.

As can bee seen in Figures 7.1-7.2 two of the four parameters that I estimate appear to have distributions close to their theoretical asymptotic distributions (show in the red dotted line). However, in all of these simulations we get a biased estimated and in general a slightly larger variance. Examining the two multinomial transition probability parameters, θ_{30} and θ_{31}, we see that the both

[9] I chose 118 months since this is the maximum duration of data that is used for estimation in Rust [9].

have a relatively large bias and a substantially larger variance than they should. This pattern also holds when the sample size increases to 13,000.

The mean and standard deviation of these distributions are shown in Table 7.4. We can see that the mean squared error decreases proportionally to the increase in sample size for the smallest sample sizes, with only a marginal decrease in the mean squared error between the 8,000 observation sample and 13,000 observation sample.

6.2 Distributional Results

Using the procedure described above, I produced simulated datasets, each containing 13,000 bus-month observations and used the nested fixed-point algorithm to estimate $\{RC, \theta_{11}, \theta_{30}, \theta_{31}\}$. Knittel and Metaxoglou [6] have shown that the choice of starting values can create very different results in highly non-linear environments. Therefore, in all of the simulations I use the same starting value, which is within 0.1 of the true value.

In order to explore the sensitivity of of the nested fixed point algorithm to the assumption that the errors are distributed Type I extreme value, I will generate datasets with errors from three different distributions: Type I extreme value, Gaussian, and a Student's T with 3 degrees of freedom. I choose these distributions since they all have unbounded support. The Gaussian distribution is useful since it is similar in shape to the Type I extreme value. Meanwhile the Student's T_3 simulation will allow me to examine the behavior of the NFXP when I have many "large" errors.

6.2.1 Extreme Value Unobservables

Of the 3,000 datasets that were created, 1,148 datasets produced results that converged using our criterion that when the gradient times the direction is less than 1×10^{-8}. The remaining datasets

produced parameter estimates where the gradient was ∞ or $-\infty$. While there are many different ways that the algorithm may be modified to get these datasets to converge, I will simply throw out these datasets from the analysis to focus on this particular procedure.

The means and standard deviations of these parameters can be seen in Table 7.5. Table 7.5 suggests that the nested fixed-point algorithm does not provide unbiased estimates of the true parameters in our sample. Two possible explanations for this naturally present themselves. First, by only using 100 buses each for 118 months, I may not have gotten close enough to ∞ to have a consistent estimate of the parameters. Secondly, my results may be biased because of the datasets that did not converge. It seems likely that these datasets may be different in some systematic way.

I reject the null hypothesis that the mean of the parameters from the simulations is equal to true mean.[10] Figures 7.5-7.8 display the full distribution of parameters from the simulations. All of the parameters appear to be distributed approximately Gaussian as the theory suggests, however formal tests, such as the Shapiro Wilk tests reject the null hypothesis that the data are Gaussian.

6.2.2 Gaussian Unobservables

Qualitatively, the simulation results from datasets that have Gaussian disturbances are similar to those with Extreme Value disturbances. There were 1,163 datasets that converged using the same convergence criterion, while the rest produced parameter estimates where the gradient was ∞ or $-\infty$.

The second panel of Table 7.5 shows descriptive statistics of these parameter. Note that the mean squared error from these results is approximately 1.5 times larger than that from the simulations with extreme value disturbances. The mean squared error for the parameters of the multino-

[10] I have not taken into account that the observations are estimated and therefore the standard errors should be larger. However, in my opinion, it is unlikely that this would change the results substantively.

mial distribution does not change much in relative terms across any of the simulations, suggesting that the likelihood function is relatively well behaved in these dimensions.

The full distribution of parameters is displayed in Figures 7.5-7.8 with an overlaid Gaussian distribution. Again, using formal tests of normality I reject the null that the estimates are distributed Gaussian.

6.2.3 Student's T_3 Unobservables

The simulation results from datasets that have Student's T_3 disturbances are quite different from the other two simulations. The center of the parameter estimates is biased substantially downward, with a mean square error of 10,000 times that from the extreme value disturbances for one parameter and 200 times larger for another parameter. These results are shown in the bottom panel of Table 7.5 with the full distribution of parameters displayed in Figures 7.5-7.8. Since the Student's T_3 distribution has has a higher probability of getting extreme values for the disturbance term, particularly extreme negative values, it makes sense that I end up with results that are biased substantially downward.

7 Conclusion

Empirical applications of highly nonlinear estimators has grown extensively recently. Naturally, these studies rely on asymptotic properties derived in the literature. However, there has been little examination of how these estimators perform in finite samples.

This paper adds to the growing literature that explores the numerical and finite sample behavior of nonlinear structural estimators. This study asks the question of how much data is "enough"

to use asymptotic results for inference about the estimated structural parameters. By simulating datasets produced knowing that the model is correctly specified, I examine the marginal distributions of parameter estimates and find them to be non-Gaussian.

The results suggest that the NFXP performs relatively poorly for samples sizes smaller than 13,000. However, there appears to be substantial gains in terms of mean squared error up to at least 8,000 observations, at which point we see the mean square error decreasing less rapidly than moving between smaller sample sizes. I can reject the null hypothesis that the empirical distribution is Gaussian for all of the parameter $\times$ sample sizes in this paper, with the distributions of the transition probabilities performing the most poorly.

One reason that the estimates may appear non-Gaussian is that the simulated datasets do not have enough observations (results are proved as either $T \to \infty$ or $N \to \infty$). However, each of these datasets contain at least 13,000 observations, which is more than many structural models have at their disposal.[11] Therefore, we should be cautious about inference that we draw from finite samples smaller than our simulation sample size.

This paper has also explored to what extent one particular estimator, the nested fixed point algorithm, depend upon distributional assumptions. Though the NFXP is rarely used due to the computational burden of computing a fixed point at every iteration, it is part of a larger class of nested-pseudo likelihood estimators that depend on distributional assumptions. We have found that when the distributional assumptions are met, the estimator performs similarly to the theory. However, as we move away from the assumed distribution, we get worse parameter estimates with a mean square error of up to 10,000 times larger than the mean square error when the assumptions

[11] Rust [9] estimates his model on 8,156 observations and Berry, Levinsohn, and Pakes [3] use 2,271 model/year observations in their seminal paper. Berry et. al do not use the nested fixed-point algorithm though their objective function is also highly nonlinear.

are met.

References

[1] Victor Aguirregabiria and Pedro Mira. Swapping the nested fixed point algorithm: A class of estimators for discrete markov decision models. *Econometrica*, 70(4):1519–1543, July 2002.

[2] Victor Aguirregabiria and Pedro Mira. Dynamic discrete choice structural models: A survey. *University of Toronto Working Paper 297*, 2007.

[3] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, July 1995.

[4] V. Joseph Hotz and Robert A. Miller. Conditional choice probabilites and the estimation of dynamic models. *Review of Economic Studies*, 60:497–529, 1993.

[5] V. Joseph Hotz, Robert A. Miller, Seth Sanders, and Jeffrey Smith. A simulation estimator for dynamic models of discrete choice. *Review of Economic Studies*, 61(2):265–289, April 1994.

[6] Christopher R Knittel and Konstantinos Metaxoglou. Estimation of random coefficient demand models: Challenges, difficulties and warnings. *Unpublished Manuscript*, 2008.

[7] Robert A. Miller. Estimating models of dynamic optimization with microeconomic data. In H. Pesaran and P. Smidth, editors, *Handbook of Applied Econometrics: Microeconomics: Volume 2*. Blackwell, 1997.

[8] Ariel Pakes. Dynamic structural models, problems and prospects. In C. Sims, editor, *Advances in Econometrics*. Cambridge University Press, 1994.

[9] John Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica*, 55(5):999–1033, September 1987.

[10] John Rust. Maximum likelihood estimation of discrete control processes. *SIAM Journal of Control and Optimization*, 26(5):1006–1024, September 1988.

[11] John Rust. Estimation of markov decision processes. In R.E. Engle and D. McFadden, editors, *Handbook of Econometrics Volume 4*. 1994.

[12] John Rust. Nested fixed point algorithm documentation manual. *Manuscript*, 2000.

[13] Wendelin Werner. Newton-like methods for the computation of fixed points. *Computational Mathematics with Applications*, 10(1):77–86, 1984.

Table 7.1: Summary of Replacement Data
(Buses where at least 1 replacement occurred)

	Mileage at Replacement				Elapsed Time (Months)				
Bus Group	Max	Min	Mean	Standard Deviation	Max	Min	Mean	Standard Deviation	Number of Observations
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	273,400	124,800	199,733	37,459	74	38	59.1	10.9	27
4	387,300	121,300	257,336	65,477	116	28	73.7	23.3	33
5	322,500	118,000	245,291	60,258	127	31	85.4	29.7	11
6	237,200	82,400	150,786	61,007	127	49	74.7	35.2	7
7	331,800	121,000	208,963	48,981	104	41	68.3	16.9	27
8	297,500	132,000	186,700	43,956	104	36	58.4	22.2	19
Full Sample	387,400	83,400	216,354	60,475	127	28	68.1	22.4	124

Source: Rust [9].

Table 7.2: Censored Data
(Subsample of buses for which no replacements occurred)

	Mileage at Replacement				Elapsed Time (Months)				
Bus Group	Max	Min	Mean	Standard Deviation	Max	Min	Mean	Standard Deviation	Number of Observations
1	120,151	65,643	100,117	12,929	25	25	25	0	15
2	161,748	142,009	151,183	8,530	49	49	49	0	4
3	280,802	199,626	250,766	21,325	75	75	75	0	21
4	352,450	310,910	337,222	17,802	118	117	117.8	0.45	5
5	326,843	326,843	326,843	0	130	130	130	0	1
6	299,040	232,395	265,264	33,332	130	128	129.3	1.15	3
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
Full Sample	352,450	65,643	207,782	85,208	130	25	66.4	34.6	49

Source: Rust [9].

Table 7.3: Structural Estimates from Rust 1987
Fixed Point Dimension = 90, $\beta = 0.9999$

Parameter	Rust Estimate	Replication Estimate
RC	9.7558	8.5075
θ_{11}	2.6275	0.7571
θ_{30}	0.3489	0.3491
θ_{31}	0.6394	0.6396
Log-Likelihood	-6055.25	-6053.55
Observations	8,156	8,156

Source: Rust [9] and author's calculations.

Table 7.4: Summary Statistics of Parameter Estimates
Model Assumptions Satisfied

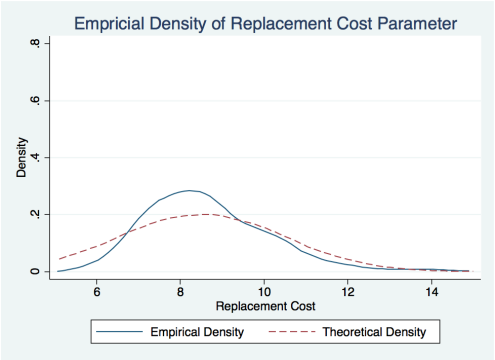
Observations	Parameter	True Mean	Mean	Standard Deviation	Mean Squared Error ($\times 10^3$)
$1,000$	RC	8.50	8.940	4.772	22962.9
	θ_1	0.76	0.862	0.561	326.2
	θ_{30}	0.35	0.334	0.023	0.7
	θ_{31}	0.64	0.614	0.038	2.1
$2,000$	RC	8.50	8.999	1.706	3149.7
	θ_1	0.76	0.880	0.360	144.8
	θ_{30}	0.35	0.335	0.020	0.6
	θ_{31}	0.64	0.617	0.033	1.6
$4,000$	RC	8.50	8.906	1.176	1541.9
	θ_1	0.76	0.846	0.263	76.9
	θ_{30}	0.35	0.336	0.018	0.5
	θ_{31}	0.64	0.617	0.032	1.5
$8,000$	RC	8.50	8.962	0.906	1025.7
	θ_1	0.76	0.847	0.202	49.1
	θ_{30}	0.35	0.336	0.017	0.5
	θ_{31}	0.64	0.618	0.031	1.4

Table 7.5: Summary Statistics of Parameter Estimates
13,000 Observations

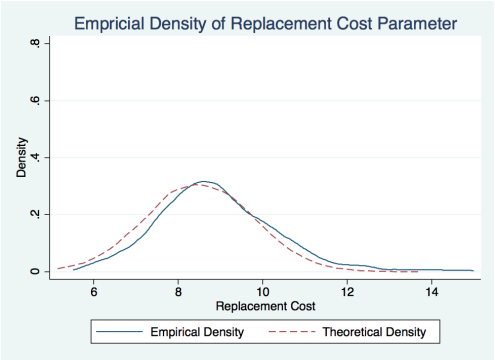
	Parameter	True Mean	Mean	Standard Deviation	Mean Squared Error ($\times 10^3$)
EV1	RC	8.50	8.903	0.725	682.0
	θ_1	0.76	0.835	0.159	31.4
	θ_{30}	0.35	0.336	0.016	0.4
	θ_{31}	0.64	0.617	0.029	1.3
Gaussian	RC	8.50	9.008	0.812	908.3
	θ_1	0.76	0.882	0.170	44.4
	θ_{30}	0.35	0.335	0.016	0.5
	θ_{31}	0.64	0.616	0.030	1.5
T_3	RC	8.50	5.875	0.286	7016.1
	θ_1	0.76	0.510	0.095	70.0
	θ_{30}	0.35	0.341	0.021	0.5
	θ_{31}	0.64	0.627	0.038	1.6

Figure 7.1: Simulations of ”Small” Data Sets: Replacement Cost Parameter

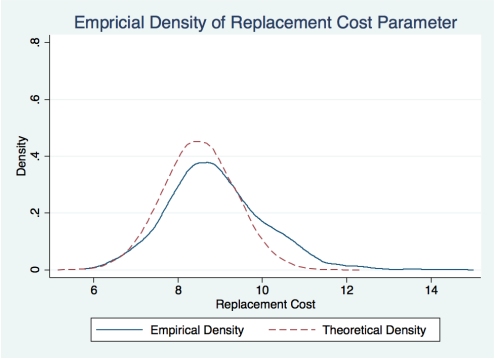
1,000 Observations



2,000 Observations



4,000 Observations



8,000 Observations

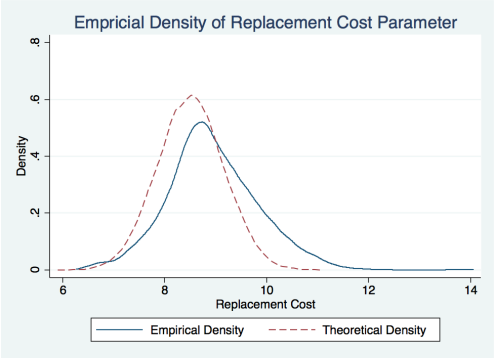
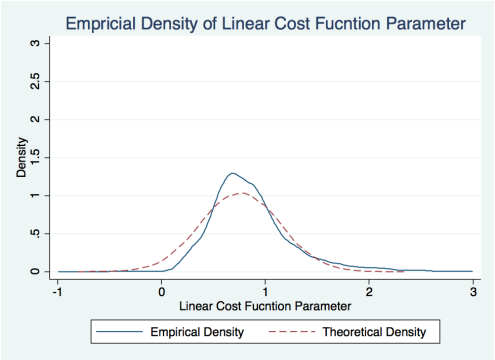
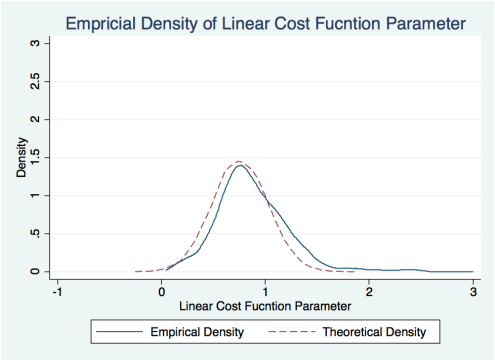


Figure 7.2: Simulations of ”Small” Data Sets: Cost Function Parameter

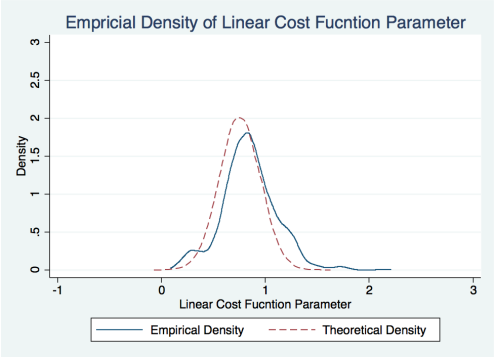
1,000 Observations



2,000 Observations



4,000 Observations



8,000 Observations

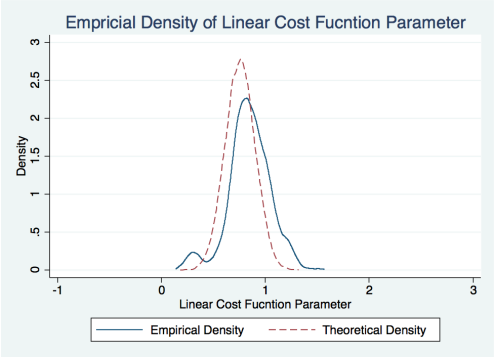
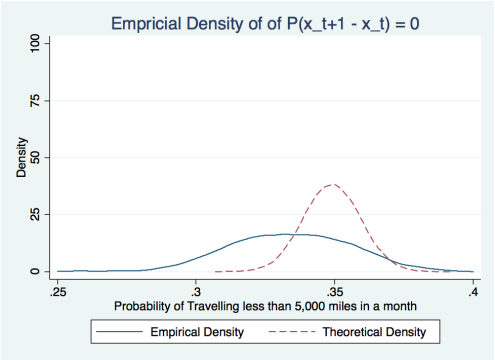
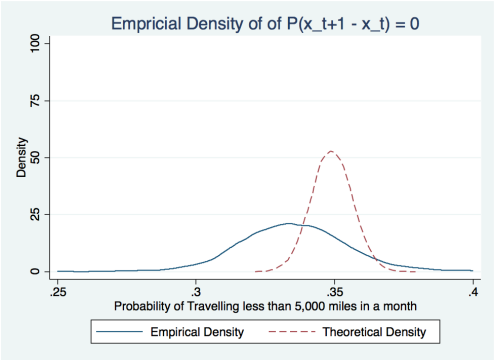


Figure 7.3: Simulations of "Small" Data Sets: $P(x_{t+1} - x_t = 0)$

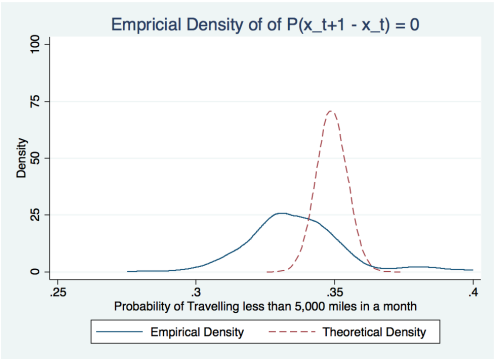
1,000 Observations



2,000 Observations



4,000 Observations



8,000 Observations

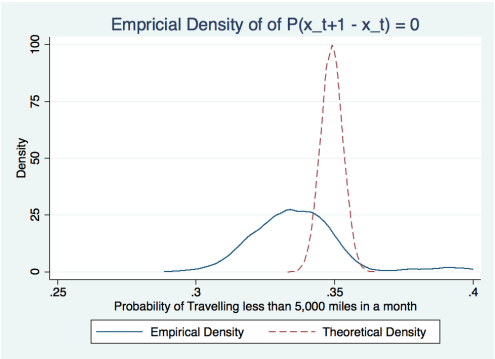
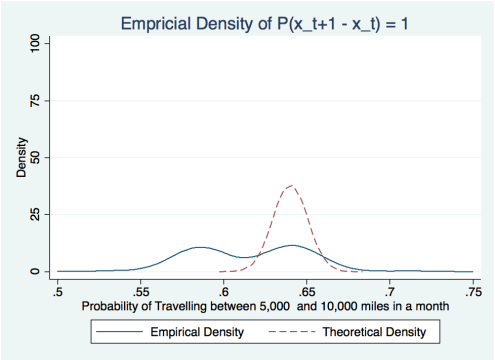
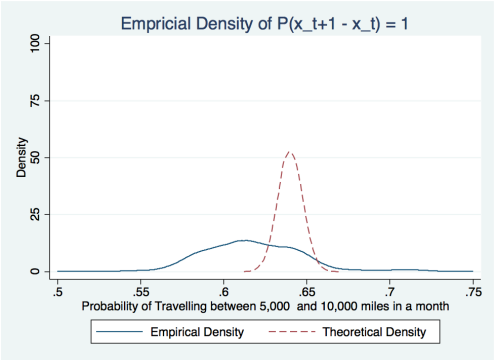


Figure 7.4: Simulations from "Small" Data Sets: $P(x_{t+1} - x_t = 1)$

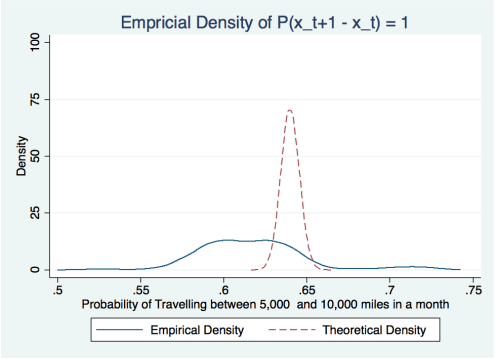
1,000 Observations



2,000 Observations



4,000 Observations



8,000 Observations

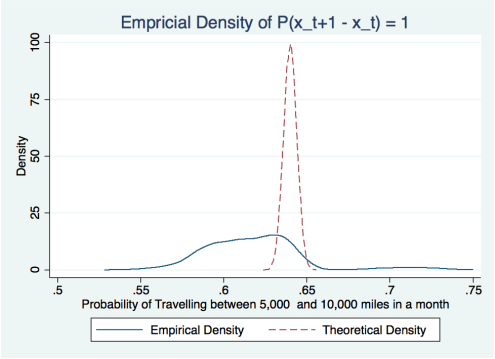
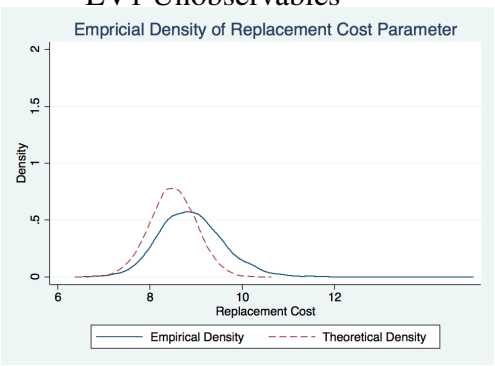
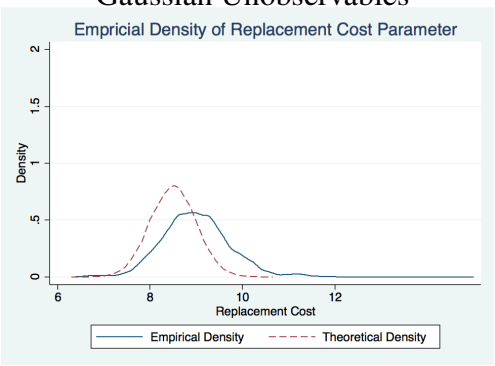


Figure 7.5: Simulations from Different Unobservable Distributions: Replacement Cost Parameter

EV1 Unobservables



Gaussian Unobservables



T_3 Unobservables

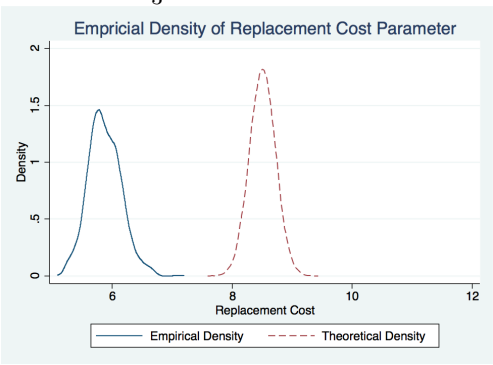
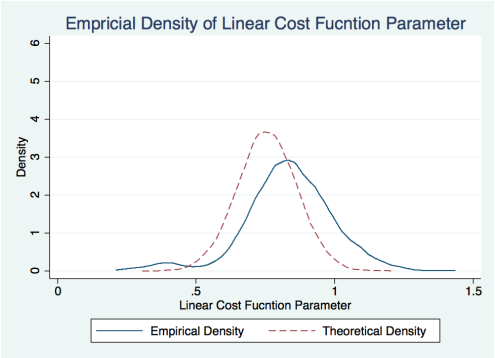
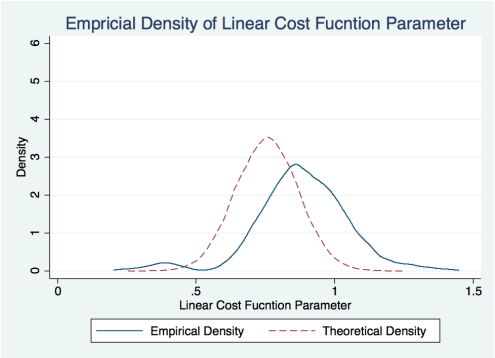


Figure 7.6: Simulations from Different Unobservable Distributions: Cost Function Parameter

EV1 Unobservables



Gaussian Unobservables



T_3 Unobservables

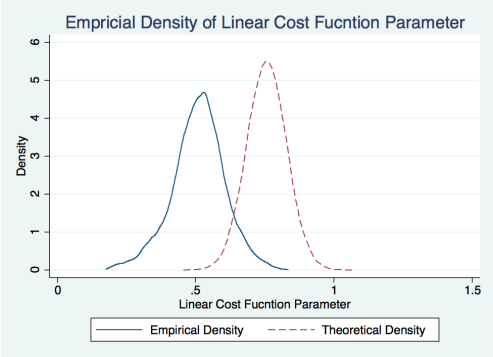
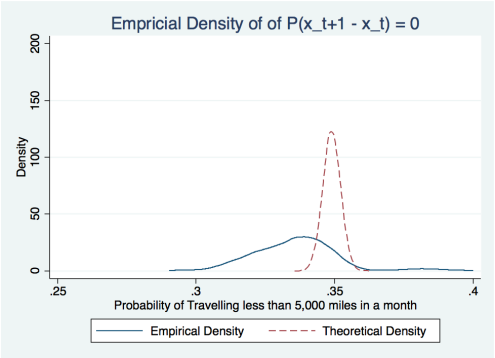
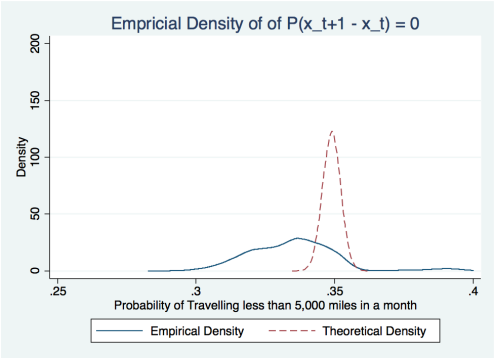


Figure 7.7: Simulations from Different Unobservable Distributions: $P(x_{t+1} - x_t = 0)$

EV1 Unobservables



Gaussian Unobservables



T_3 Unobservables

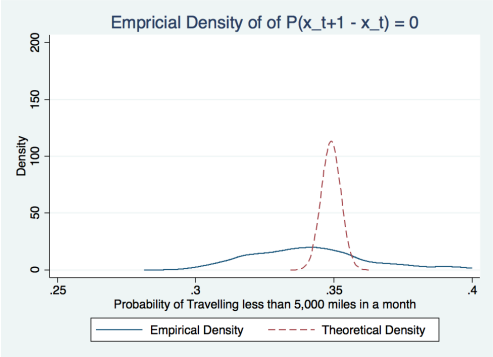
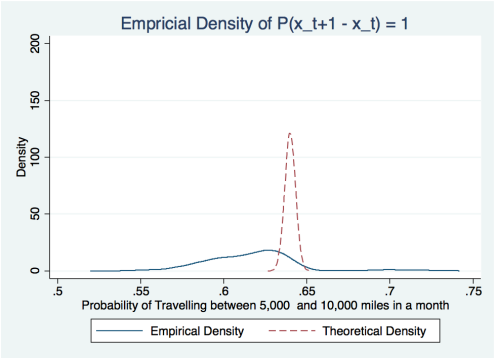
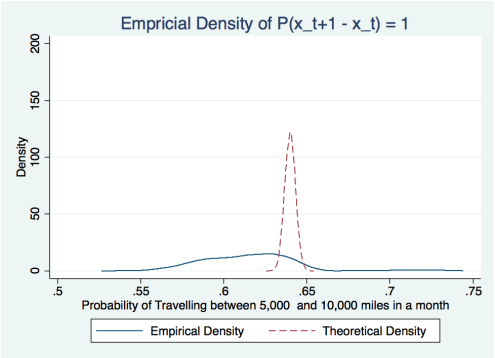


Figure 7.8: Simulations from Different Unobservable Distributions: $P(x_{t+1} - x_t = 1)$

EV1 Unobservables



Gaussian Unobservables



T_3 Unobservables